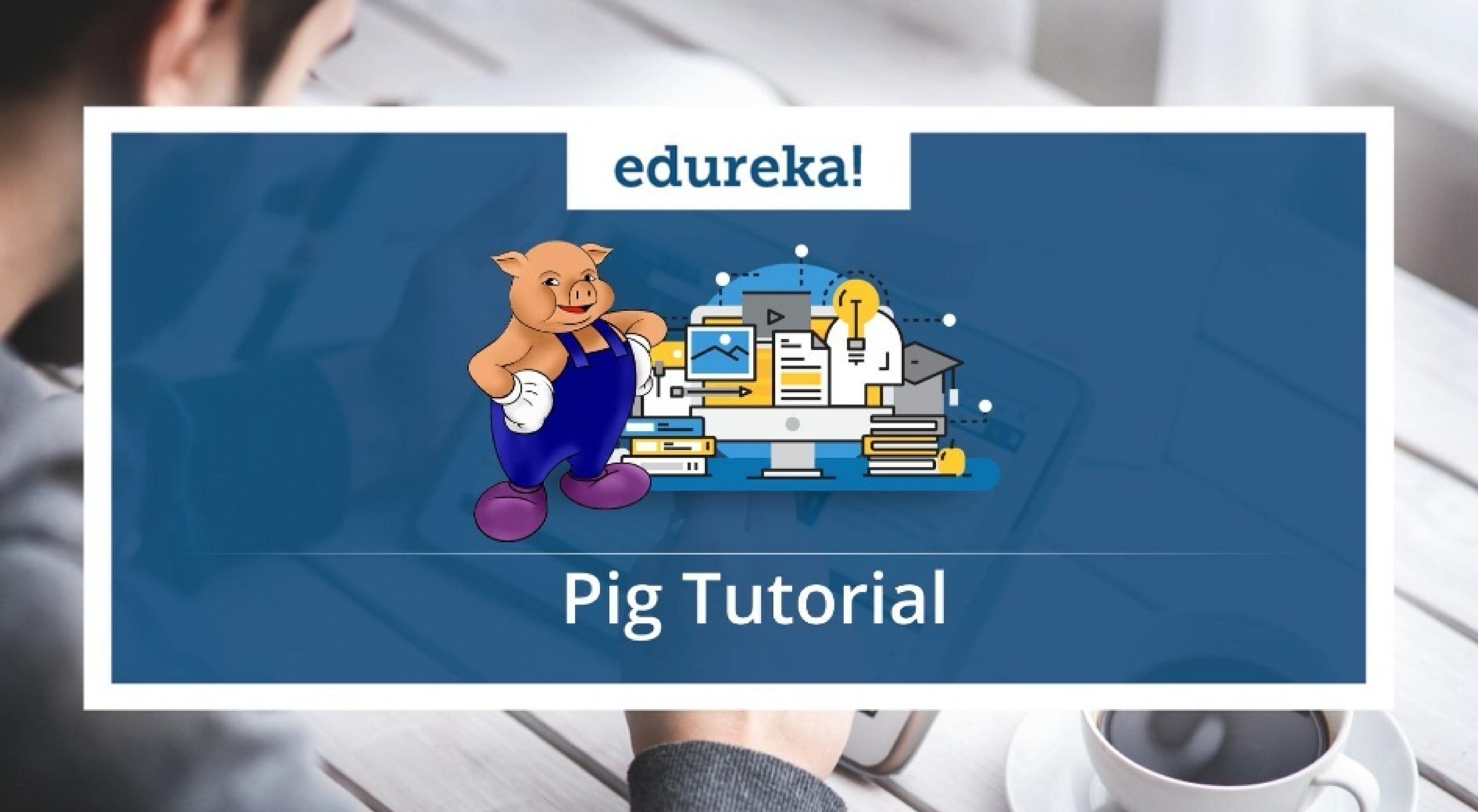

edureka!
Pig Tutorial

Agenda for today's Session

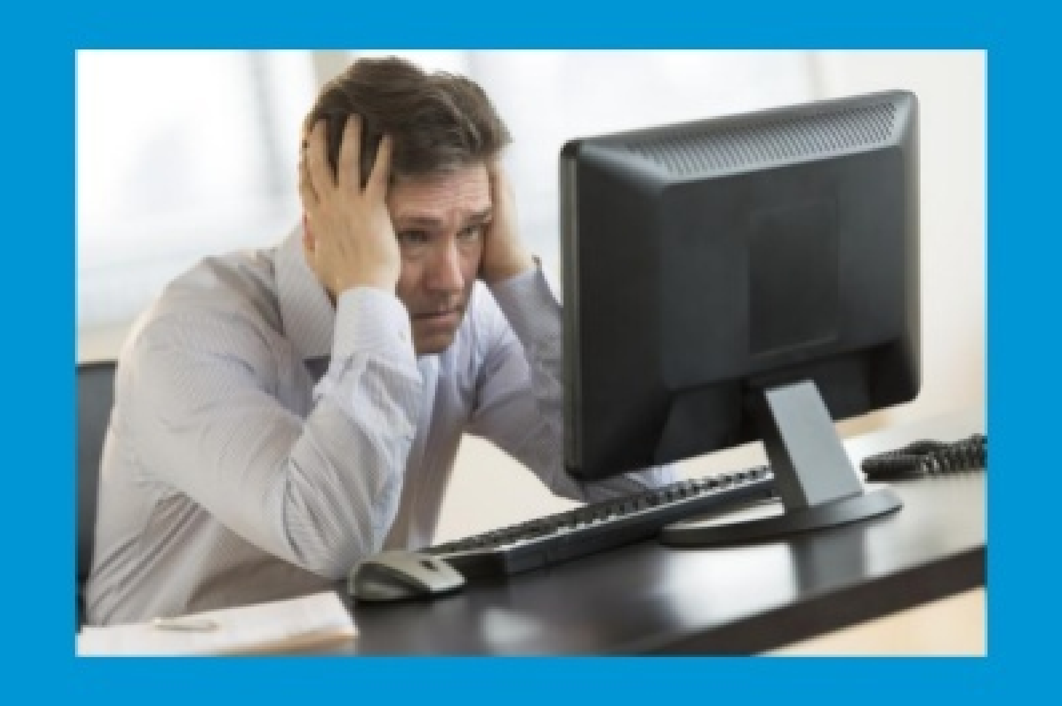
- Entry of Apache Pig
- Pig vs MapReduce
- Twitter Case Study on Apache Pig
- Apache Pig Architecture
- Pig Components
- Pig Data Model & Operators
- Running Pig Commands and Pig Scripts (Log Analysis)



MapReduce Way



In MapReduce, you need to write a program in Java/Python to process the data.



What if you are from Non-programming background!!

Are your Hadoop days over before they even started? ☹



No need to worry at all!

There are multiple tools in Hadoop Ecosystem where you do not need programming background.

And in today's session, I will tell you about one such tool!

Apache PIG

edureka!

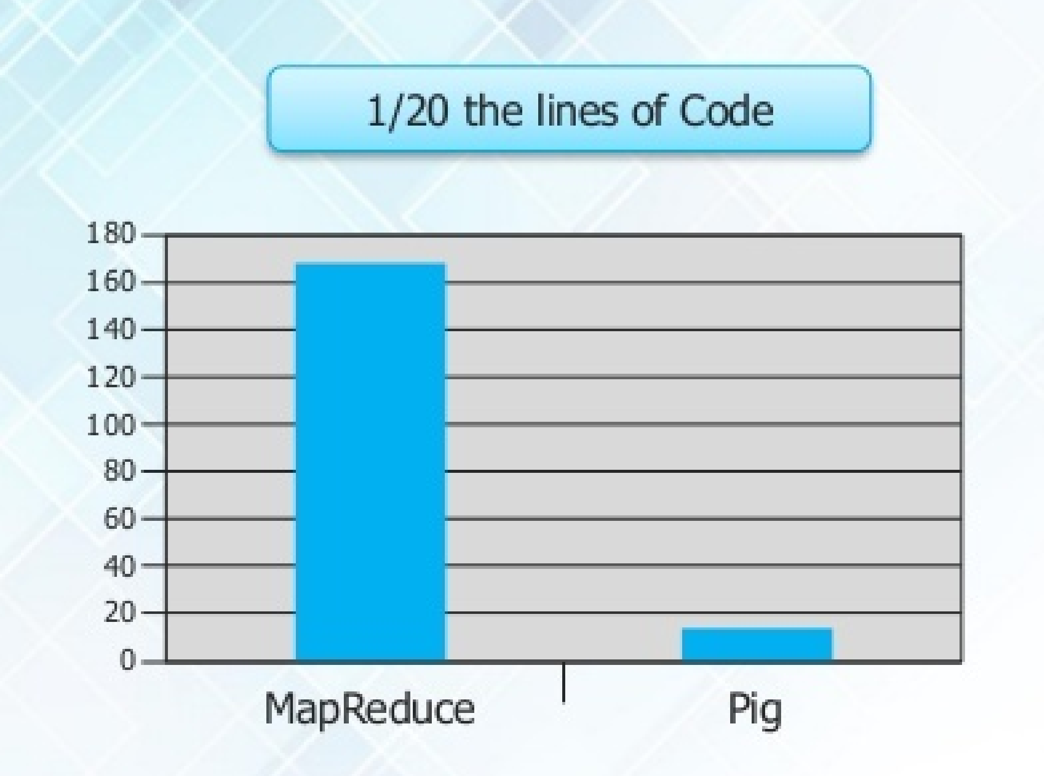


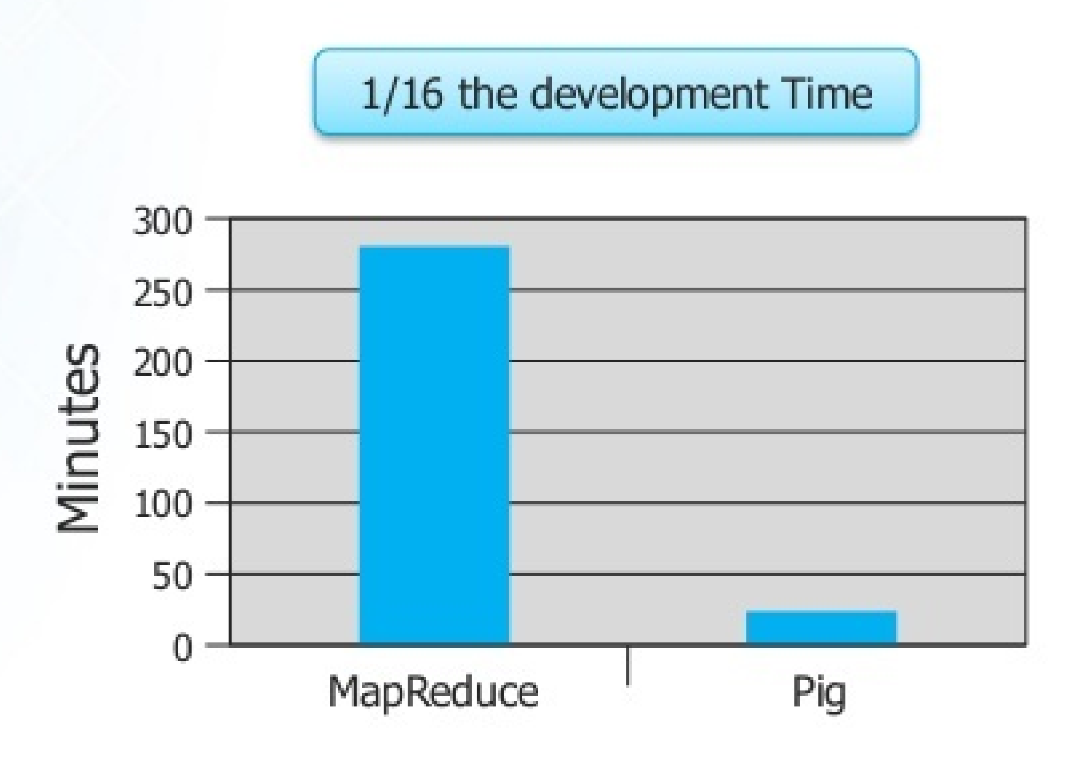
- An open-source high-level dataflow system
- Introduced by Yahoo
- Provides abstraction over MapReduce
- Two main components – the *Pig Latin* language and the *Pig Execution*

Fun Fact:
✓ ***10 lines of pig latin= approx. 200 lines of Map-Reduce Java Program***

Why go for PIG when MR is there?

edureka!





Apache Pig vs MapReduce

Apache Pig vs MapReduce

edureka!



- High-level data flow tool
- No need to write complex programs
- Built-in support for data operations like joins, filters, ordering, sorting etc.
- Provides nested data types like tuples, bags, and maps

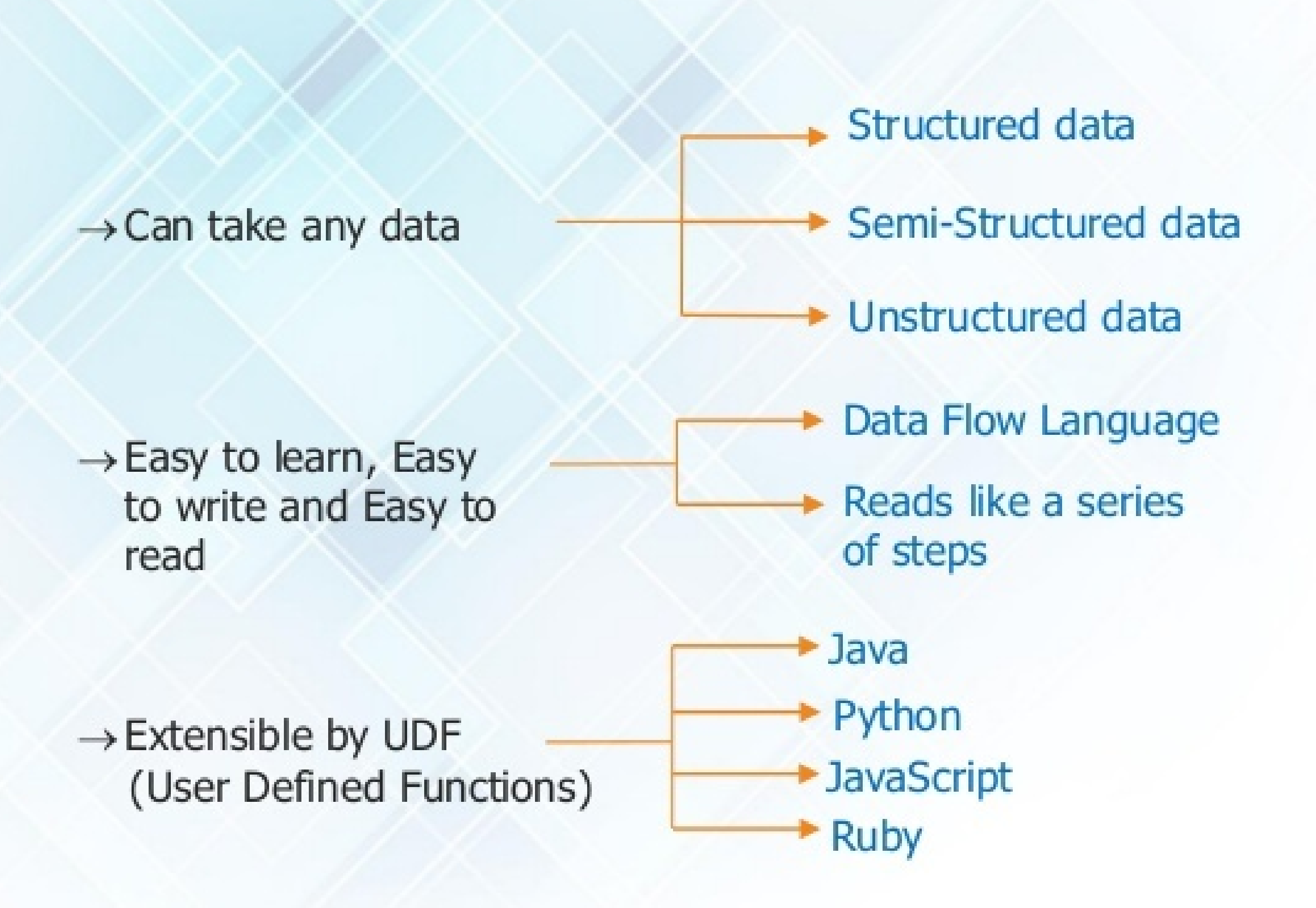


- Low-level data processing paradigm
- You need write programs in Java/Python etc.
- Performing data operations in MapReduce is a humongous task
- Nested data types are not there in MapReduce

Some more reasons to choose Apache Pig

Why Apache Pig?

edureka!



→ Provides common data operations filters, joins, ordering, etc. and nested data types tuples, bags, and maps missing from MapReduce.

→ An ad-hoc way of creating and executing map-reduce jobs on very large data sets

→ Open source and actively supported by a community of developers.

Twitter Case Study

Twitter Case Study

edureka!

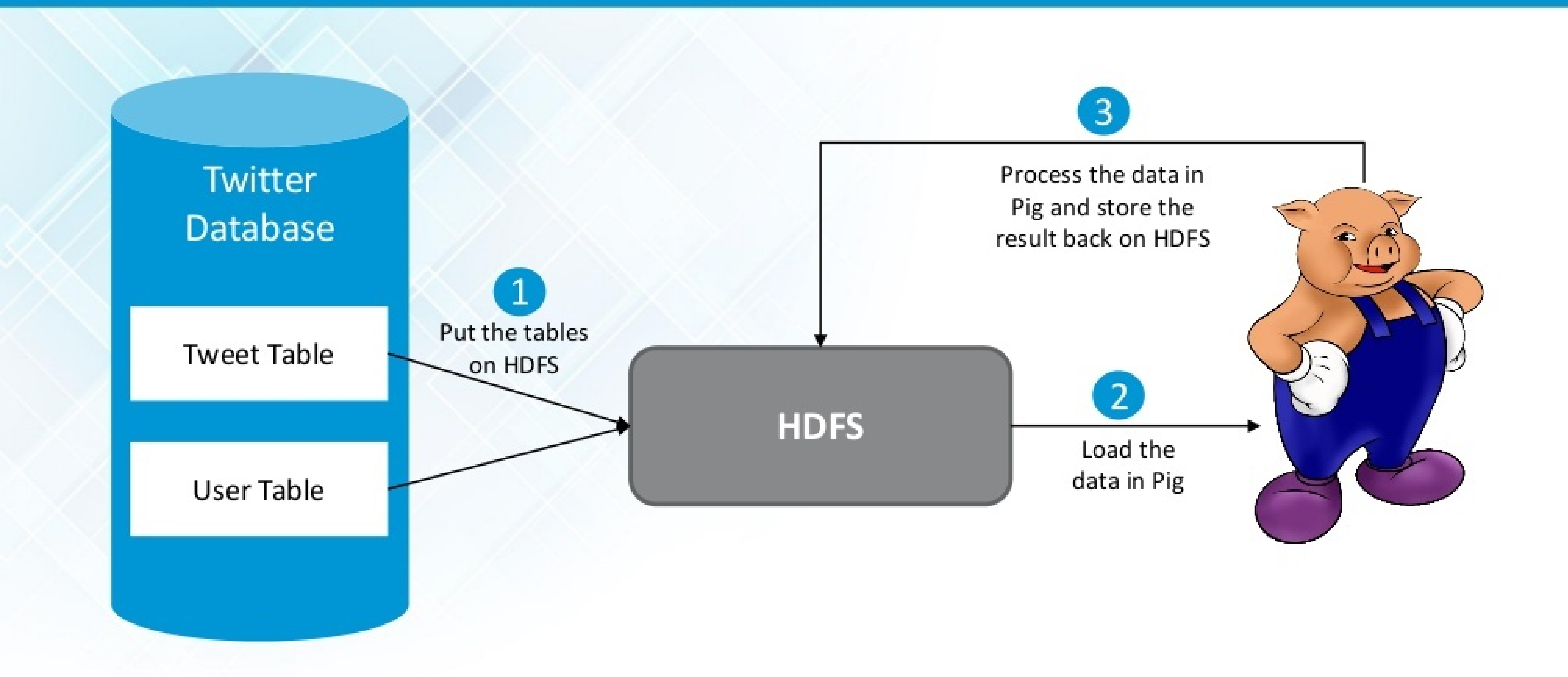


- Twitter's data was growing at an accelerating rate (i.e. 10 TB/day).
- Thus, Twitter decided to move the archived data to HDFS and adopt Hadoop for extracting the business values out of it.
- Their major aim was to analyse data stored in Hadoop to come up with the multiple insights on a daily, weekly or monthly basis.

Let me talk about one of the insight they wanted to know.

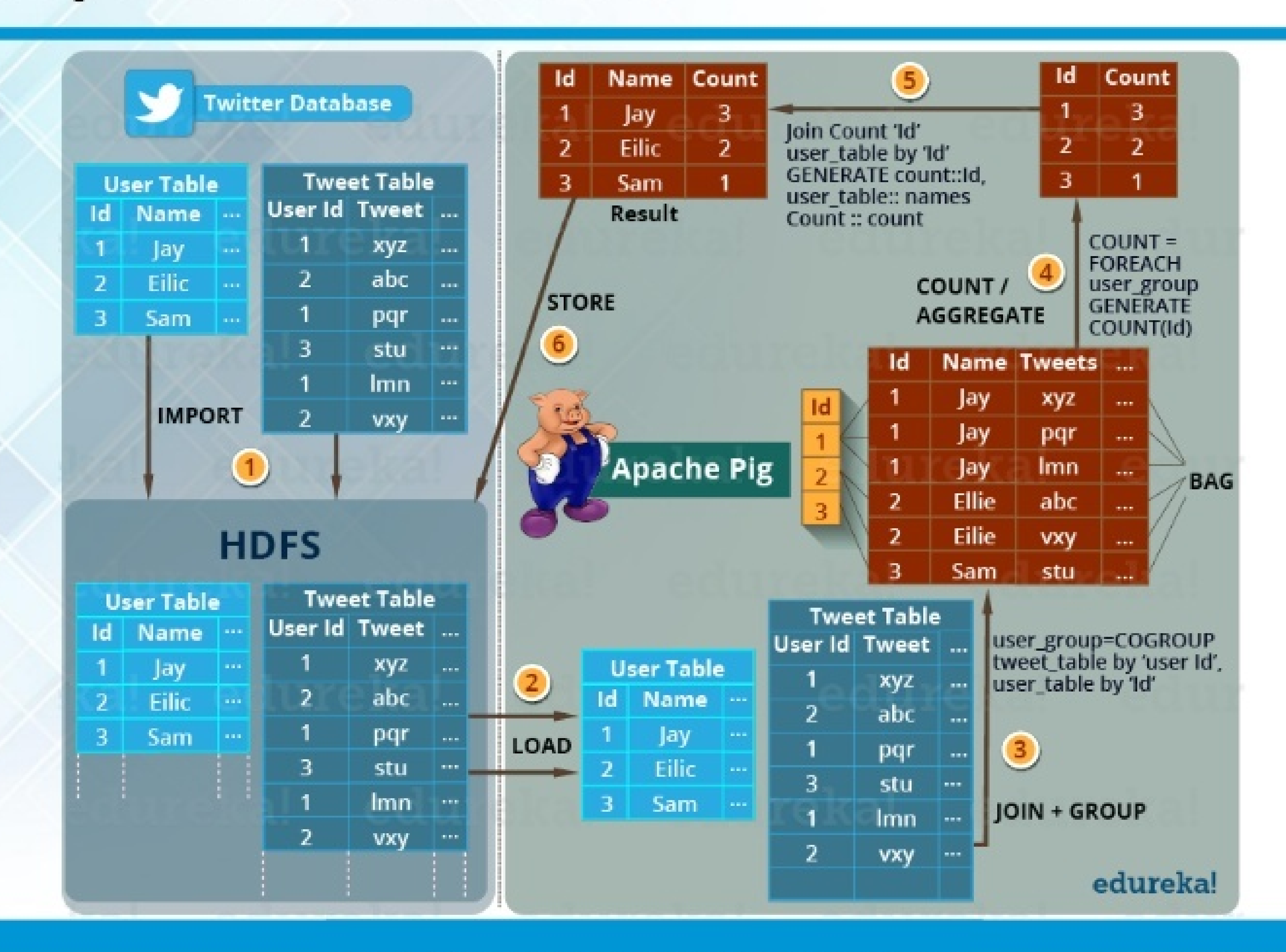
Analyzing how many tweets are stored per user, in the given tweet tables?

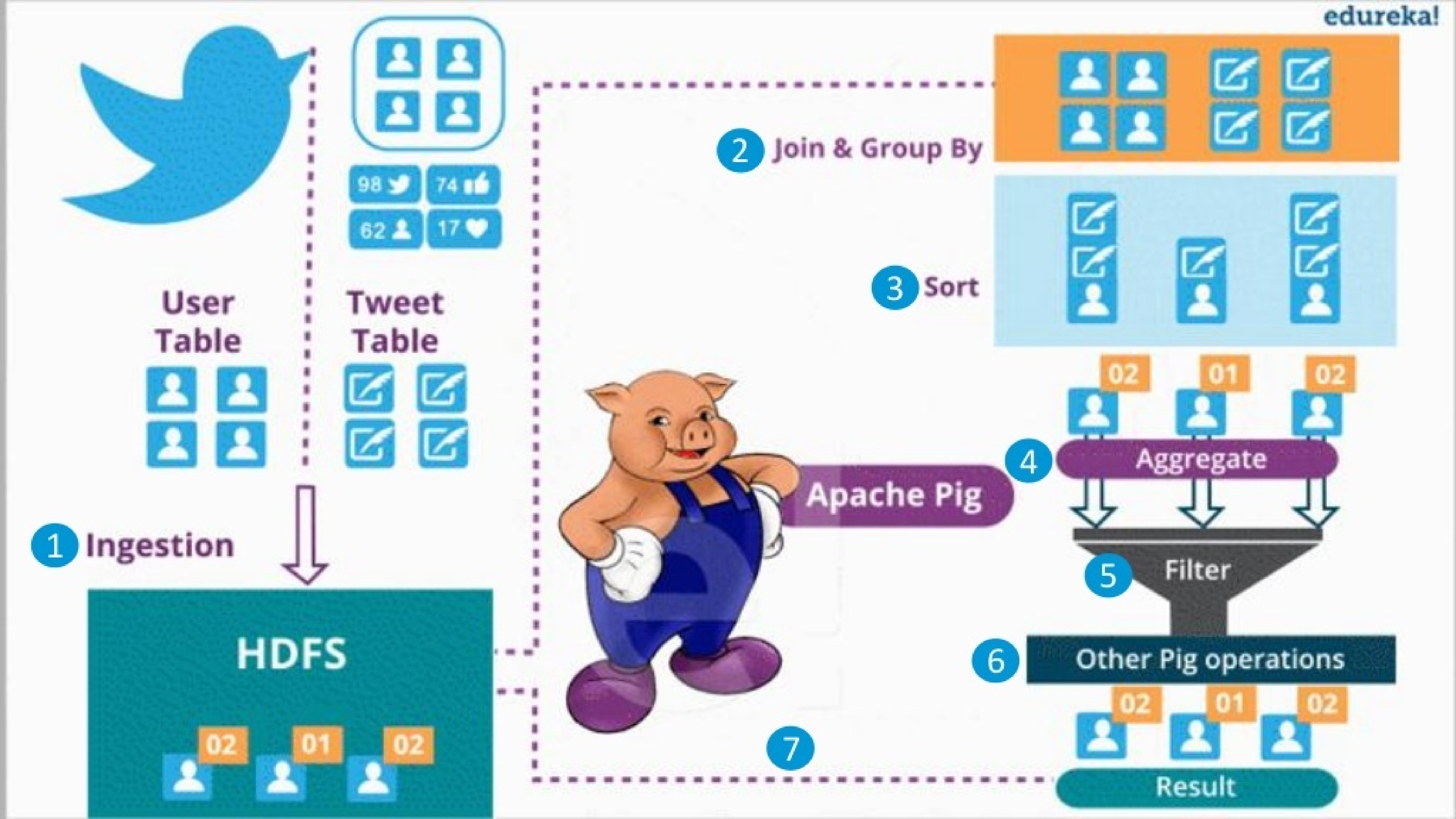
High Level Implementation



Detailed Implementation Flow

edureka!





edureka!
User Table
Tweet Table
98
74
62
17
1 Ingestion
HDFS
02
01
02
2 Join & Group By
3 Sort
02
01
02
4 Aggregate
Apache Pig
5 Filter
6 Other Pig operations
02
01
02
7
Result

Apache Pig Architecture

Apache Pig Architecture

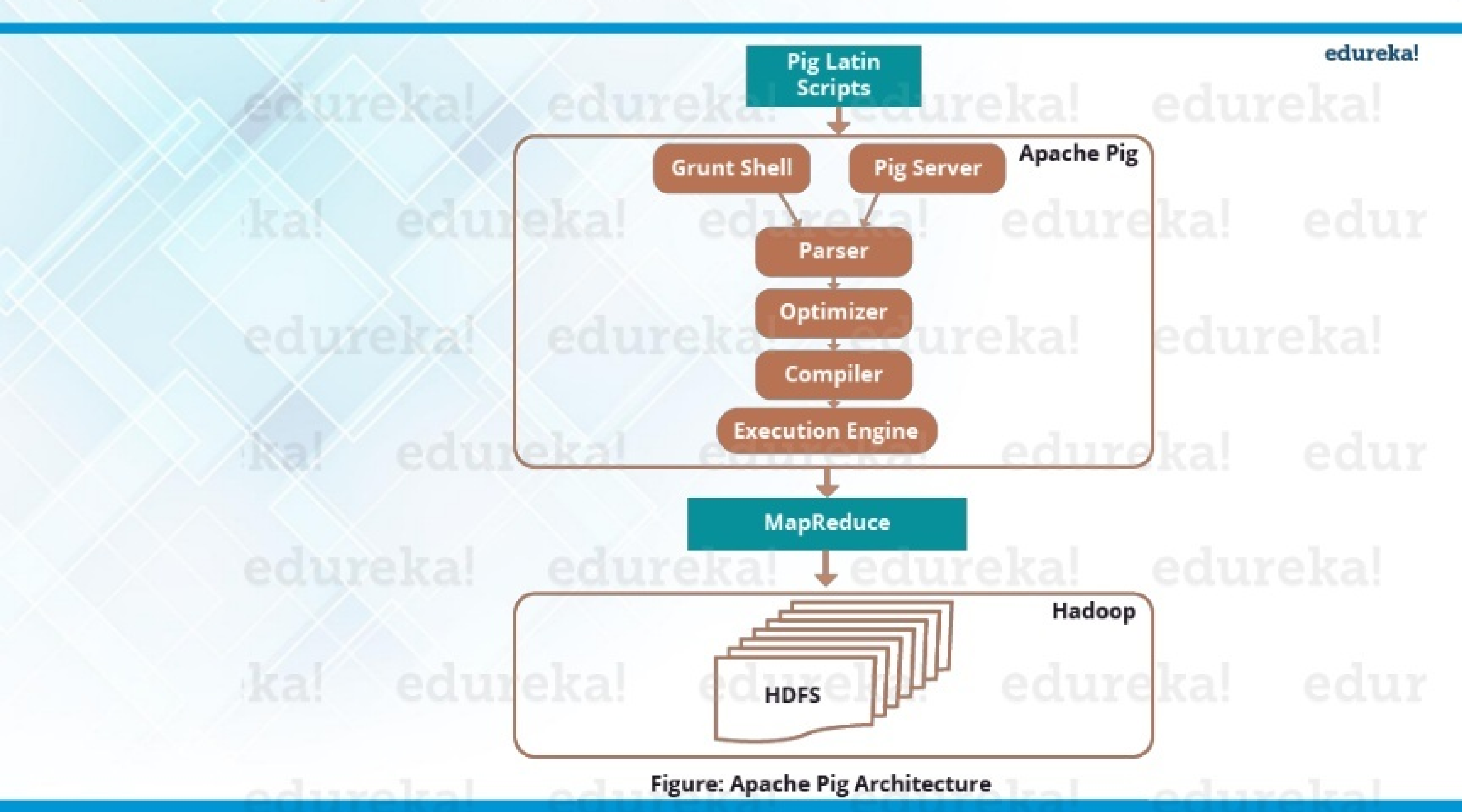
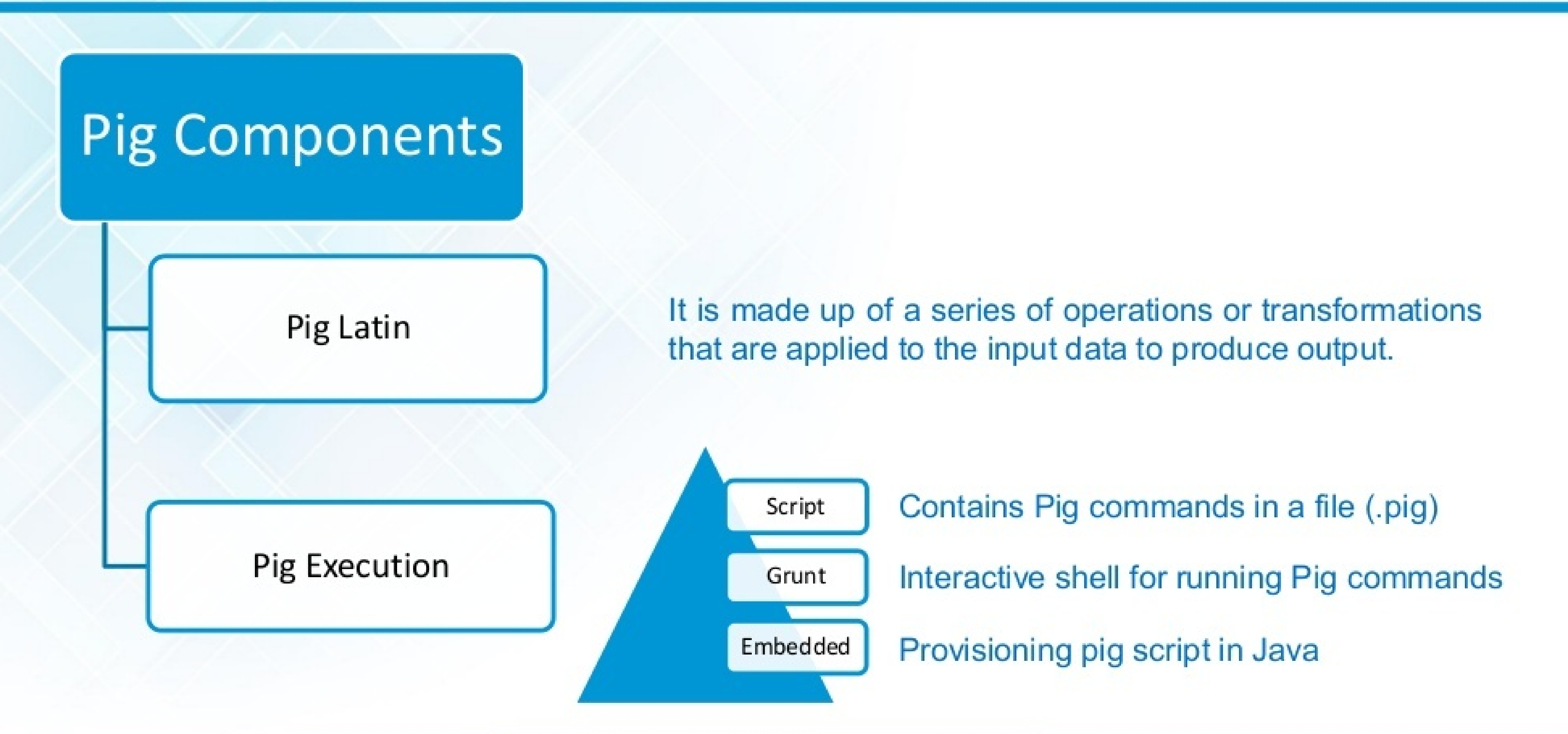


Figure: Apache Pig Architecture

Apache Pig Components

Apache Pig Components

edureka!



Apache Pig Running Modes

Pig Running Modes

You can run Apache Pig in 2 modes:

MapReduce Mode – This is the default mode, which requires access to a Hadoop cluster and HDFS installation. The input and output in this mode are present on HDFS.

Command: pig

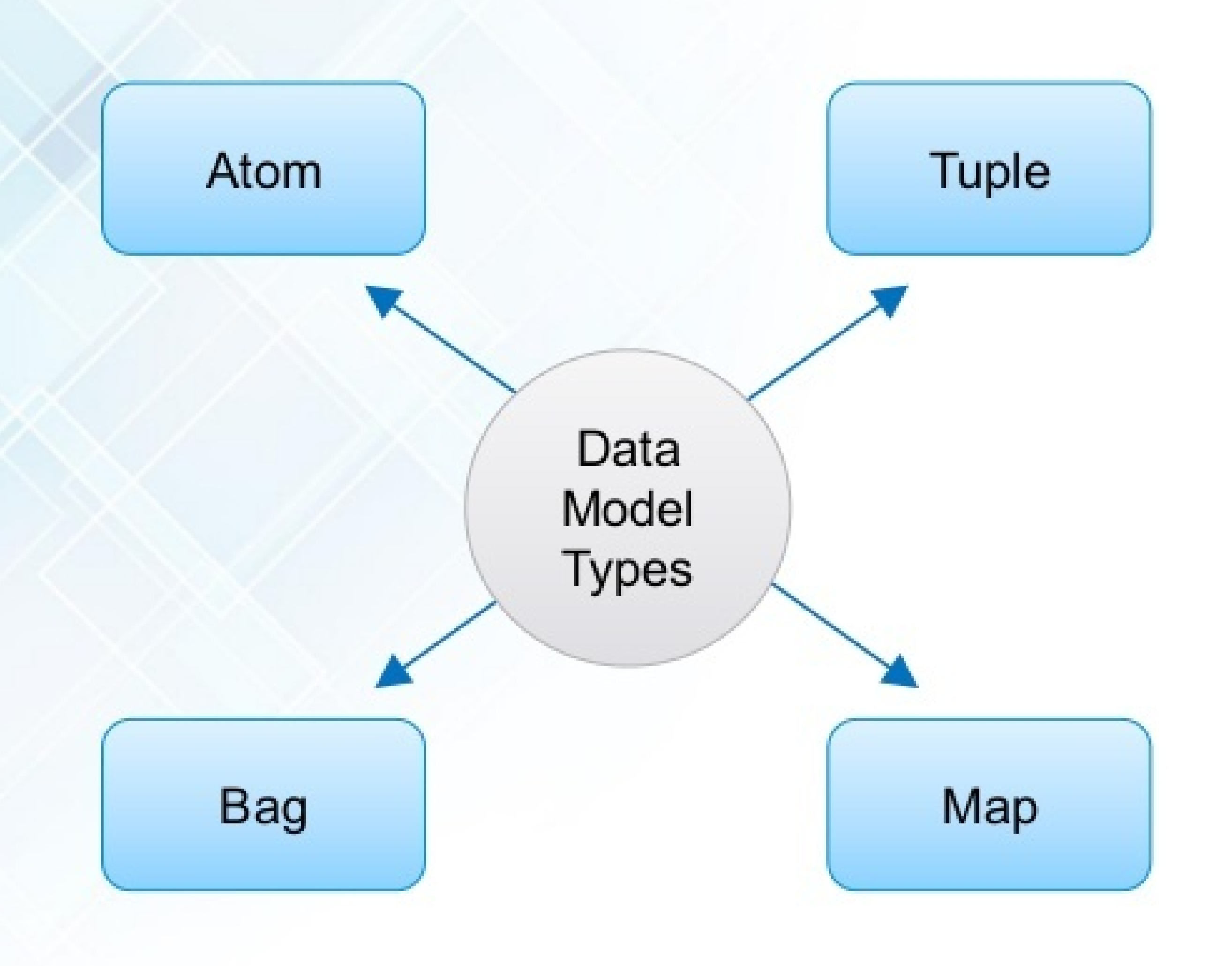
Local Mode – With access to a single machine, all files are installed and run using a local host and file system. Here the local mode is specified using '-x flag' (pig -x local). The input and output in this mode are present on local file system.

Command: pig –x local

Before going to practical, let us understand Data Models in Pig

Pig Data Model

edureka!



Pig Data Model – Tuple and Bag

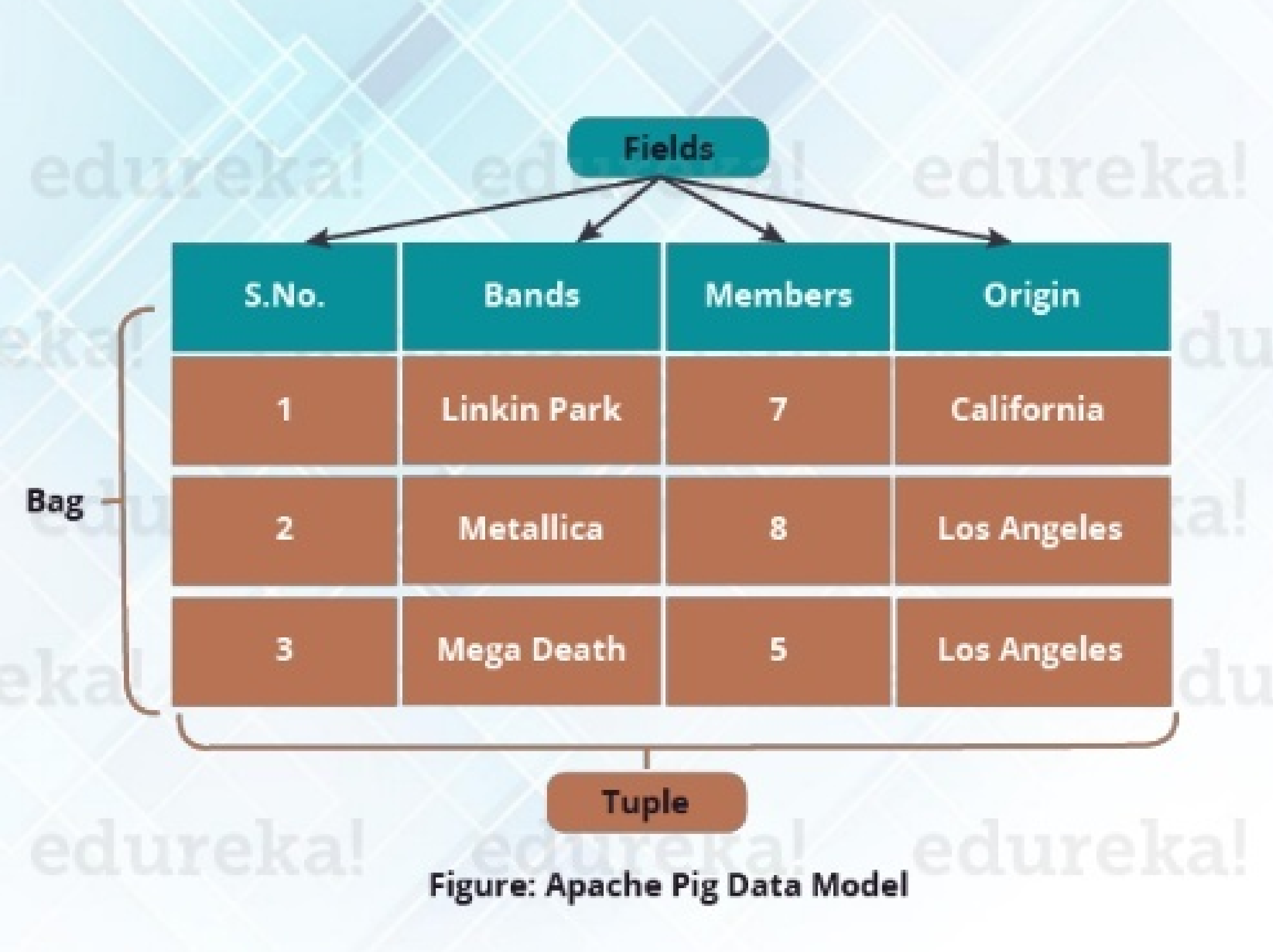


Figure: Apache Pig Data Model

- **Tuple** is an ordered set of fields which may contain different data types for each field.

  Example of tuple − (1, Linkin Park, 7, California)

- A **Bag** is a collection of a set of tuples and these tuples are subset of rows or entire rows of a table.

  Example of a bag − {(Linkin Park, 7, California), (Metallica, 8), (Mega Death, Los Angeles)}

Pig Data Model – Map and Atom



Keys

Maps

Band # Linkin Park, Member # 7, Origin # California

Band # Metallica, Member # 8, Origin # Los Angeles

Band # Mega Death, Member # 8, Origin # Los Angeles

Values

Figure: Map Example

- A Map is key-value pairs used to represent data elements.

  Example of maps– [band#Linkin Park, members#7], [band#Metallica, members#8]

- Atoms are basic data types which are used in all the languages like string, int, float, long, double, char[], byte[]

Pig Operators

Pig Operators

edureka!

Operator	Description
LOAD	Load data from the local file system or HDFS storage into Pig
FOREACH	Generates data transformations based on columns of data
FILTER	Selects tuples from a relation based on a condition
JOIN	Join the relations based on the column
ORDER BY	Sort a relation based on one or more fields
STORE	Save results to the local file system or HDFS
DISTINCT	Removes duplicate tuples in a relation
GROUP	Groups together the tuples with the same group key (key field)
COGROUP	It is same as GROUP. But COGROUP is used when multiple relations re involved

Let us execute few Pig commands on grunt shell

Analysing Logs Using Apache Pig

Analysing Logs Using Apache Pig

edureka!

- There is an application which processes sampleclass recordings.
- Here is a log file which is recording all the events happening when the application is running.
- We will analyse this log file to understand what are the types of event happening in this log file and the count of each event.

sample.log (~) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

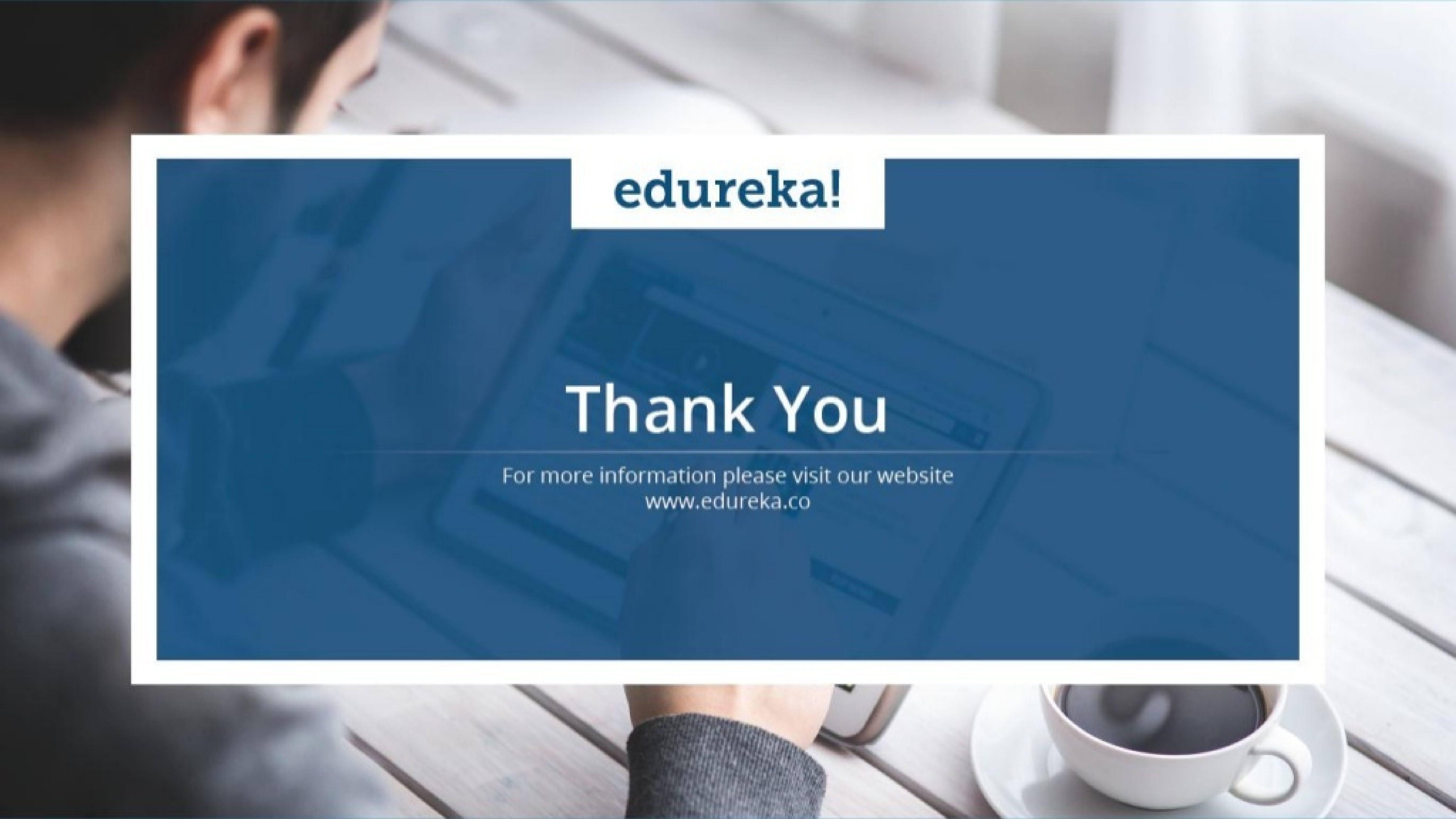
sample.log

```
2012-02-03 18:35:34 SampleClass6 [INFO] everything normal for id 577725851
2012-02-03 18:35:34 SampleClass4 [FATAL] system problem at id 1991281254
2012-02-03 18:35:34 SampleClass3 [DEBUG] detail for id 1304807656
2012-02-03 18:35:34 SampleClass3 [WARN] missing id 423340895
2012-02-03 18:35:34 SampleClass5 [TRACE] verbose detail for id 2082654978
2012-02-03 18:35:34 SampleClass0 [ERROR] incorrect id  1886438513
2012-02-03 18:35:34 SampleClass9 [TRACE] verbose detail for id 438634209
2012-02-03 18:35:34 SampleClass8 [DEBUG] detail for id 2074121310
2012-02-03 18:35:34 SampleClass0 [TRACE] verbose detail for id 1505582508
2012-02-03 18:35:34 SampleClass0 [TRACE] verbose detail for id 1903854437
2012-02-03 18:35:34 SampleClass7 [DEBUG] detail for id 915853141
2012-02-03 18:35:34 SampleClass3 [TRACE] verbose detail for id 303132401
```

Create and Run a Pig Script to Analyze the logs

Learning Resources

- **Hadoop Tutorial:** www.edureka.co/blog/hadoop-tutorial
- **Pig Tutorial:** https://www.edureka.co/blog/pig-tutorial
- **Operators in Pig:** https://www.edureka.co/blog/operators-in-apache-pig/


edureka!
Thank You
For more information please visit our website
www.edureka.co